Logistic Regression: From Linear to Classification

Nipun Batra

IIT Gandhinagar

July 30, 2025

Outline

1. From Regression to Classification
2. The Logistic Function
3. Odds and Log-Odds
4. Maximum Likelihood Estimation
5. Summary
6. Deriving Cost Function via Maximum Likelihood Estimation

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

Issues with linear regression:

- Predictions can be < 0 or > 1 (not valid probabilities!)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

Issues with linear regression:

- Predictions can be < 0 or > 1 (not valid probabilities!)
- Outliers heavily influence the decision boundary

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

Issues with linear regression:

- Predictions can be < 0 or > 1 (not valid probabilities!)
- Outliers heavily influence the decision boundary
- Assumes constant variance (inappropriate for binary outcomes)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

Issues with linear regression:

- Predictions can be < 0 or > 1 (not valid probabilities!)
- Outliers heavily influence the decision boundary
- Assumes constant variance (inappropriate for binary outcomes)

The Problem with Linear Regression for Classification

Question: Can we use linear regression for classification?

Example: Email Spam Detection

- Input: Email features (word counts, length, etc.)
- Output: $y = 1$ (spam) or $y = 0$ (not spam)

Issues with linear regression:

- Predictions can be < 0 or > 1 (not valid probabilities!)
- Outliers heavily influence the decision boundary
- Assumes constant variance (inappropriate for binary outcomes)

What We Really Want

Predict **probabilities** $P(y = 1|x)$ where $0 \leq P \leq 1$

Pop Quiz: Why Not Linear Regression?

Quick Quiz 1

What's wrong with using linear regression $\hat{y} = \mathbf{X}\boldsymbol{\theta}$ for binary classification?

a) It's too slow to compute

Answer: b) Linear regression can predict negative values or values > 1, which aren't valid probabilities!

Pop Quiz: Why Not Linear Regression?

Quick Quiz 1

What's wrong with using linear regression $\hat{y} = \mathbf{X}\boldsymbol{\theta}$ for binary classification?

a) It's too slow to compute
b) Predictions can be outside [0,1] range

Answer: b) Linear regression can predict negative values or values > 1, which aren't valid probabilities!

Pop Quiz: Why Not Linear Regression?

Quick Quiz 1

What's wrong with using linear regression $\hat{y} = \mathbf{X}\boldsymbol{\theta}$ for binary classification?

a) It's too slow to compute
b) Predictions can be outside [0,1] range
c) It only works for numerical features

Answer: b) Linear regression can predict negative values or values > 1, which aren't valid probabilities!

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)
- $z \to -\infty \implies \sigma(z) \to 0$ (high confidence: negative class)

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)
- $z \to -\infty \implies \sigma(z) \to 0$ (high confidence: negative class)
- $z = 0 \implies \sigma(z) = 0.5$ (uncertain: decision boundary)

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)
- $z \to -\infty \implies \sigma(z) \to 0$ (high confidence: negative class)
- $z = 0 \implies \sigma(z) = 0.5$ (uncertain: decision boundary)
- Always outputs values in $(0, 1)$ ✓

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)
- $z \to -\infty \implies \sigma(z) \to 0$ (high confidence: negative class)
- $z = 0 \implies \sigma(z) = 0.5$ (uncertain: decision boundary)
- Always outputs values in $(0, 1)$ ✓

The Sigmoid/Logistic Function

Solution: Transform linear output to probability using the **sigmoid function**

Logistic Regression Model

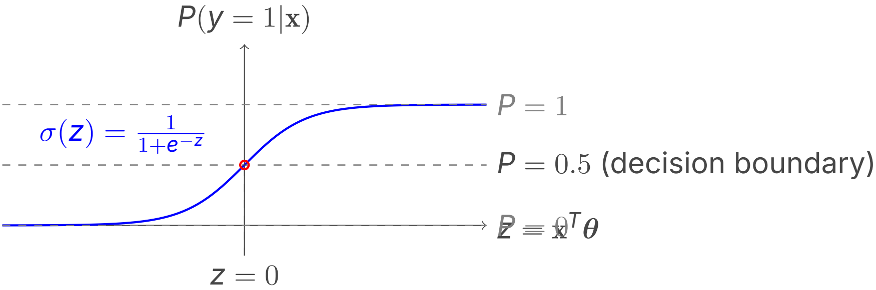
$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}}$$

Key properties of sigmoid $\sigma(z)$:

- $z \to +\infty \implies \sigma(z) \to 1$ (high confidence: positive class)
- $z \to -\infty \implies \sigma(z) \to 0$ (high confidence: negative class)
- $z = 0 \implies \sigma(z) = 0.5$ (uncertain: decision boundary)
- Always outputs values in $(0, 1)$ ✓

Interpretation

Visualizing the Sigmoid Function



Decision rule: Predict class 1 if $P(y = 1|\mathbf{x}) > 0.5$, else class 0

Pop Quiz: Sigmoid Properties

Quick Quiz 2

What is the value of the sigmoid function at $z = 0$?

a) $\sigma(0) = 0$

Answer: b) $\sigma(0) = \frac{1}{1+e^{-0}} = \frac{1}{1+1} = 0.5$ (decision boundary)

Pop Quiz: Sigmoid Properties

Quick Quiz 2

What is the value of the sigmoid function at $z = 0$?

a) $\sigma(0) = 0$

b) $\sigma(0) = 0.5$

Answer: b) $\sigma(0) = \frac{1}{1+e^{-0}} = \frac{1}{1+1} = 0.5$ (decision boundary)

Pop Quiz: Sigmoid Properties

Quick Quiz 2

What is the value of the sigmoid function at $z = 0$?

a) $\sigma(0) = 0$
b) $\sigma(0) = 0.5$
c) $\sigma(0) = 1$

Answer: b) $\sigma(0) = \frac{1}{1+e^{-0}} = \frac{1}{1+1} = 0.5$ (decision boundary)

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Odds Ratio

$$\text{Odds} = \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})} = e^{\mathbf{x}^T\boldsymbol{\theta}}$$

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Odds Ratio

$$\text{Odds} = \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})} = e^{\mathbf{x}^T\boldsymbol{\theta}}$$

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1 + e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Odds Ratio

$$\text{Odds} = \frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} = \frac{P(y = 1|\mathbf{x})}{1 - P(y = 1|\mathbf{x})} = e^{\mathbf{x}^T\boldsymbol{\theta}}$$

Log-Odds (Logit)

$$\log(\text{Odds}) = \mathbf{x}^T\boldsymbol{\theta} = \log\frac{P(y = 1|\mathbf{x})}{1 - P(y = 1|\mathbf{x})}$$

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Odds Ratio

$$\text{Odds} = \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})} = e^{\mathbf{x}^T\boldsymbol{\theta}}$$

Log-Odds (Logit)

$$\log(\text{Odds}) = \mathbf{x}^T\boldsymbol{\theta} = \log\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}$$

Understanding the Log-Odds

From sigmoid to interpretable quantities:

$$P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x}) = \frac{e^{-\mathbf{x}^T\boldsymbol{\theta}}}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}} \tag{1}$$

Odds Ratio

$$\text{Odds} = \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})} = e^{\mathbf{x}^T\boldsymbol{\theta}}$$

Log-Odds (Logit)

$$\log(\text{Odds}) = \mathbf{x}^T\boldsymbol{\theta} = \log\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}$$

Interpretation: Linear model predicts log-odds, not

Why Not Squared Loss?

Problem: Squared loss + sigmoid creates non-convex optimization

- Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ is non-linear

Why Not Squared Loss?

Problem: Squared loss + sigmoid creates non-convex optimization

- Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ is non-linear
- Composition $(\sigma(\mathbf{x}^T\boldsymbol{\theta}) - y)^2$ has multiple local minima

Why Not Squared Loss?

Problem: Squared loss + sigmoid creates non-convex optimization

- Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ is non-linear
- Composition $(\sigma(\mathbf{x}^T\boldsymbol{\theta}) - y)^2$ has multiple local minima
- No guarantee gradient descent finds global optimum

Why Not Squared Loss?

Problem: Squared loss + sigmoid creates non-convex optimization

- Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ is non-linear
- Composition $(\sigma(\mathbf{x}^T\boldsymbol{\theta}) - y)^2$ has multiple local minima
- No guarantee gradient descent finds global optimum

Why Not Squared Loss?

Problem: Squared loss + sigmoid creates non-convex optimization

- Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ is non-linear
- Composition $(\sigma(\mathbf{x}^T\boldsymbol{\theta}) - y)^2$ has multiple local minima
- No guarantee gradient descent finds global optimum

Better Approach: Maximum Likelihood

Design a loss function that creates a **convex** optimization problem

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = P(y = 1|\mathbf{x})^y \cdot P(y = 0|\mathbf{x})^{1-y} = p^y(1 - p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = P(y = 1|\mathbf{x})^y \cdot P(y = 0|\mathbf{x})^{1-y} = p^y(1 - p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y \log p + (1 - y) \log(1 - p)$$

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = P(y = 1|\mathbf{x})^y \cdot P(y = 0|\mathbf{x})^{1-y} = p^y(1-p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y \log p + (1-y)\log(1-p)$$

Cross-Entropy Loss (Negative Log-Likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = -y \log p - (1-y)\log(1-p)$$
$$= -y \log \sigma(\mathbf{x}^T\boldsymbol{\theta}) - (1-y)\log(1-\sigma(\mathbf{x}^T\boldsymbol{\theta}))$$

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x},\boldsymbol{\theta}) = P(y=1|\mathbf{x})^y \cdot P(y=0|\mathbf{x})^{1-y} = p^y(1-p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y\log p + (1-y)\log(1-p)$$

Cross-Entropy Loss (Negative Log-Likelihood)

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= -y\log p - (1-y)\log(1-p) \\ &= -y\log\sigma(\mathbf{x}^T\boldsymbol{\theta}) - (1-y)\log(1-\sigma(\mathbf{x}^T\boldsymbol{\theta}))\end{aligned}$$

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = P(y = 1|\mathbf{x})^y \cdot P(y = 0|\mathbf{x})^{1-y} = p^y(1 - p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y \log p + (1 - y) \log(1 - p)$$

Cross-Entropy Loss (Negative Log-Likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = -y \log p - (1 - y) \log(1 - p)$$
$$= -y \log \sigma(\mathbf{x}^T\boldsymbol{\theta}) - (1 - y) \log(1 - \sigma(\mathbf{x}^T\boldsymbol{\theta}))$$

Key properties:

- Convex in θ ✓

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = P(y = 1|\mathbf{x})^y \cdot P(y = 0|\mathbf{x})^{1-y} = p^y(1 - p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y \log p + (1 - y) \log(1 - p)$$

Cross-Entropy Loss (Negative Log-Likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = -y \log p - (1 - y) \log(1 - p)$$
$$= -y \log \sigma(\mathbf{x}^T\boldsymbol{\theta}) - (1 - y) \log(1 - \sigma(\mathbf{x}^T\boldsymbol{\theta}))$$

Key properties:

- Convex in θ ✓
- Penalizes wrong predictions heavily

Cross-Entropy Loss

Likelihood for one sample:

$$P(y|\mathbf{x},\boldsymbol{\theta}) = P(y=1|\mathbf{x})^y \cdot P(y=0|\mathbf{x})^{1-y} = p^y(1-p)^{1-y}$$

where $p = \sigma(\mathbf{x}^T\boldsymbol{\theta})$

Log-likelihood:

$$\ell(\boldsymbol{\theta}) = y\log p + (1-y)\log(1-p)$$

Cross-Entropy Loss (Negative Log-Likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = -y\log p - (1-y)\log(1-p)$$
$$= -y\log\sigma(\mathbf{x}^T\boldsymbol{\theta}) - (1-y)\log(1-\sigma(\mathbf{x}^T\boldsymbol{\theta}))$$

Key properties:

- Convex in θ ✓
- Penalizes wrong predictions heavily
- Natural choice for binary classification

Pop Quiz: Cross-Entropy

Quick Quiz 3

For a true positive example ($y = 1$), what happens to the cross-entropy loss as $p \to 0$?

a) Loss approaches 0

Answer: b) Loss $= -\log p \to +\infty$ as $p \to 0$ (heavily penalizes confident wrong predictions!)

Pop Quiz: Cross-Entropy

Quick Quiz 3

For a true positive example ($y = 1$), what happens to the cross-entropy loss as $p \to 0$?

a) Loss approaches 0
b) Loss approaches $+\infty$

Answer: b) Loss $= -\log p \to +\infty$ as $p \to 0$ (heavily penalizes confident wrong predictions!)

Pop Quiz: Cross-Entropy

Quick Quiz 3

For a true positive example ($y = 1$), what happens to the cross-entropy loss as $p \to 0$?

a) Loss approaches 0
b) Loss approaches $+\infty$
c) Loss remains constant

Answer: b) Loss $= -\log p \to +\infty$ as $p \to 0$ (heavily penalizes confident wrong predictions!)

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$
- **Training:** Cross-entropy loss ensures convex optimization

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$
- **Training:** Cross-entropy loss ensures convex optimization
  - Better than squared loss for classification

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$
- **Training:** Cross-entropy loss ensures convex optimization
  - Better than squared loss for classification
  - Solvable with gradient descent

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$
- **Training:** Cross-entropy loss ensures convex optimization
  - Better than squared loss for classification
  - Solvable with gradient descent

Logistic Regression: Key Takeaways

- **Problem:** Linear regression inadequate for classification
- **Solution:** Sigmoid function maps linear output to probabilities
  - $P(y = 1|\mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\theta}}}$
- **Interpretation:** Linear model predicts log-odds
  - $\mathbf{x}^T\boldsymbol{\theta} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$
- **Training:** Cross-entropy loss ensures convex optimization
  - Better than squared loss for classification
  - Solvable with gradient descent
- **Extensions:** Multinomial logistic regression for multi-class problems

This cost function is called cross-entropy.

This cost function is called cross-entropy.
Why?

What is the interpretation of the cost function?

What is the interpretation of the cost function?
Let us try to write the cost function for a single example:

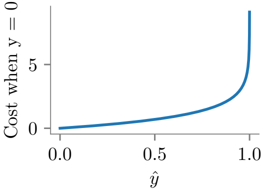
What is the interpretation of the cost function?
Let us try to write the cost function for a single example:

$$J(\theta) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

What is the interpretation of the cost function?
Let us try to write the cost function for a single example:

$$J(\theta) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

First, assume y_i is 0, then if $\hat{y}_i$ is 0, the loss is 0; but, if $\hat{y}_i$ is 1, the loss tends towards infinity!



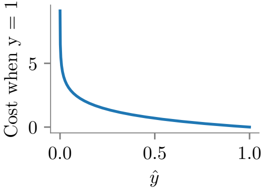
What is the interpretation of the cost function?

$$J(\theta) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

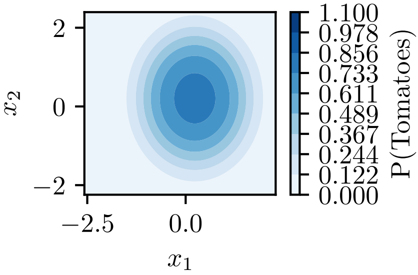
What is the interpretation of the cost function?

$$J(\theta) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

Now, assume y_i is 1, then if $\hat{y}_i$ is 0, the loss is huge; but, if $\hat{y}_i$ is 1, the loss is zero!



Bias!



2
0
−2
−2.5
0.0
x_1
x_2
1.100
0.978
0.856
0.733
0.611
0.489
0.367
0.244
0.122
0.000
P(Tomatoes)

How would you learn a classifier? Or, how would you expect the classifier to learn decision boundaries?

1. Use one-vs.-all on Binary Logistic Regression

1. Use one-vs.-all on <u>Binary</u> Logistic Regression
2. Use one-vs.-one on <u>Binary</u> Logistic Regression

1. Use one-vs.-all on Binary Logistic Regression
2. Use one-vs.-one on Binary Logistic Regression
3. Extend Binary Logistic Regression to Multi-Class Logistic Regression

1. Learn P(setosa (class 1)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_1)$

1. Learn P(setosa (class 1)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_1)$
2. P(versicolor (class 2)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_2)$

1. Learn P(setosa (class 1)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_1)$
2. P(versicolor (class 2)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_2)$
3. P(virginica (class 3)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_3)$

1. Learn P(setosa (class 1)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_1)$
2. P(versicolor (class 2)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_2)$
3. P(virginica (class 3)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_3)$
4. Goal: Learn $\theta_i \forall i \in \{1, 2, 3\}$

1. Learn P(setosa (class 1)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_1)$
2. P(versicolor (class 2)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_2)$
3. P(virginica (class 3)) = $\mathcal{F}(\mathbf{X}\boldsymbol{\theta}_3)$
4. Goal: Learn $\theta_i \forall i \in \{1, 2, 3\}$
5. Question: What could be an $\mathcal{F}$?

1. Question: What could be an $\mathcal{F}$?

1. Question: What could be an $\mathcal{F}$?
2. Property: $\sum_{i=1}^{3} \mathcal{F}(\mathbf{X}\boldsymbol{\theta}_i) = 1$

1. Question: What could be an $\mathcal{F}$?
2. Property: $\sum_{i=1}^{3} \mathcal{F}(\mathbf{X}\boldsymbol{\theta}_i) = 1$
3. Also $\mathcal{F}(z) \in [0, 1]$

1. Question: What could be an $\mathcal{F}$?
2. Property: $\sum_{i=1}^{3} \mathcal{F}(\mathbf{X}\boldsymbol{\theta}_i) = 1$
3. Also $\mathcal{F}(z) \in [0, 1]$
4. Also, $\mathcal{F}(z)$ has squashing proprties: $R \mapsto [0, 1]$

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$
= $-(0 \times \log(0.1) + 1 \times \log(0.8) + 0 \times \log(0.1))$

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$
= $-(0 \times \log(0.1) + 1 \times \log(0.8) + 0 \times \log(0.1))$
Tends to zero

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$
= $-(0 \times \log(0.1) + 1 \times \log(0.4) + 0 \times \log(0.1))$

Let us calculate $-\sum_{k=1}^{3} y_i^k \log \hat{y}_i^k$
= $-(0 \times \log(0.1) + 1 \times \log(0.4) + 0 \times \log(0.1))$
High number! Huge penalty for misclassification!

More generally,

More generally,

$$J(\theta) = -\left\{ \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right\}$$

More generally,

$$J(\theta) = -\left\{ \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right\}$$

$$J(\theta) = -\left\{ \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right\}$$

Extend to K-class:

$$J(\theta) = -\left\{ \sum_{i=1}^{N} \sum_{k=1}^{K} y_i^k \log(\hat{y}_i^k) \right\}$$

What is the key difference between sigmoid and softmax functions?

What is the key difference between sigmoid and softmax functions?

What is the key difference between sigmoid and softmax functions?

Why do we use cross-entropy loss instead of squared error?

What is the key difference between sigmoid and softmax functions?

Why do we use cross-entropy loss instead of squared error?

What is the key difference between sigmoid and softmax functions?

Why do we use cross-entropy loss instead of squared error?

How does regularization help in logistic regression?

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods
- **Cross-Entropy Loss**: Appropriate for classification problems

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods
- **Cross-Entropy Loss**: Appropriate for classification problems

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods
- **Cross-Entropy Loss**: Appropriate for classification problems
- **No Closed Form**: Requires iterative optimization (gradient descent)

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods
- **Cross-Entropy Loss**: Appropriate for classification problems
- **No Closed Form**: Requires iterative optimization (gradient descent)

Key Takeaways

- **Probabilistic Model**: Outputs probabilities via sigmoid function
- **Linear Decision Boundary**: Creates linear separation in feature space
- **Maximum Likelihood**: Optimized using gradient-based methods
- **Cross-Entropy Loss**: Appropriate for classification problems
- **No Closed Form**: Requires iterative optimization (gradient descent)
- **Regularization**: L1/L2 help prevent overfitting